Assessing Quality in Automated Metadata Aggregation and Mapping Services

Martin Höffernig, Thomas Orgel, Silvia Russegger, Werner Bailer

JOANNEUM RESEARCH Forschungsgesellschaft mbH – DIGITAL
Steyrergasse 17, 8010 Graz, Austria
{firstname.lastname}@joanneum.at

Abstract. Over the last decade, Europe has put tremendous effort into making cultural, educational and scientific resources available via online services. It is now possible to tap into this vast amount of resources and build applications making use of scientific and cultural heritage data, aggregating and converting them on the fly with services deployed in the cloud. This paper addresses the problem assessing the quality of the source metadata from different data providers and the quality of automatic mappings needed for the aggregation of heterogeneous data. We describe tools for integrated quality assessment and present the results achieved.

1 Introduction

Memory institutions, such as libraries, archives and museums host collections consisting of very different kinds of objects and archival material. These materials are used within the context of an organisation (maybe with various departments) but it becomes more and more important to spread the cultural content to a variety of professionals or to the interested public. Over the last decade, Europe has put tremendous effort into making cultural, educational and scientific resources available digitally. Based on national aggregators, initiatives like Europeana[1] nowadays provide a plethora of cultural resources for people worldwide. In parallel, the Semantic Web and the availability of Linked Open Data (LOD) have been growing exponentially, providing semantically enhanced access to and interchange of relevant scientific and cultural resources.

The EEXCESS project[2] aims at bringing such resources to a wider audience by recommending them to users in the context of their daily online activities, such as viewing web pages, writing blog posts or using educational resources. Based on their user profile and context in the activity, the EEXCESS framework defines queries, which are then federated to a set of data providers, i.e., Europeana and a range of other cultural heritage portals. One tool using this framework is the EEXCESS Browser Plugin, which is a publicly available Chrome plugin to get recommendations from participating data providers related to the

[1] http://www.europeana.eu

[2] http://www.eexcess.eu

currently viewed web page. The recommended objects are presented in a sidebar with a set of metadata and images. As users' interests and context may be very specific, it is expected that the tools can tap into the long tail of cultural heritage resources, i.e., not just provide a small set of widely consumed content, but a lot of "niche" content. Thus there is no preloaded or cached content set. Instead, the entire process happens on the fly. In order to present and rank objects from different collections, the framework needs to map the providers' metadata into a homogeneous form, requiring a mapping process that can run automatically once configured. Measures have to be taken along the online process to ensure the quality of the heterogeneous set of results returned to the consumer. This includes documenting the provenance of metadata and assessing the source metadata quality, as well as the quality of the applied mapping process. The necessary quality assessment tools need to be integrated into this process.

This paper discusses the quality assessment problem in a setting such as the EEXCESS project as well as existing work (remainder of this section). The proposed approaches for assessing source metadata quality and mapping quality and the results achieved are described in Sections 2 and 3 respectively. Section 4 concludes the paper.

In order to ensure the quality of the provided metadata, we need to address both the *source metadata quality*, i.e., the quality of records returned from a particular data provider in response to a query, and the *mapping quality*, i.e., the completeness and fidelity of the metadata in the target common data model. Assessing the source metadata quality needs to be done when connecting new data providers, but also as a background task in order to monitor the running system. Most of the existing literature on metadata quality considers the metadata of single or multiple records of a collection, i.e., our source metadata. The authors of [2] define the following measures for quality: completeness, accuracy, provenance, conformance, logical consistency and coherence, timeliness and accessibility. A taxonomy of 22 measures for information quality has been proposed in [5], grouped into three categories: intrinsic, relational/contextual and reputational information quality. Many of these aspects can only be checked manually by experts, and some are only applicable to homogeneous collections dealing with similar and related objects. An approach that attempts automation has been proposed in [1]. The authors start from viewing metadata quality as the fitness for use for a specific purpose and propose three metrics: completeness (the element is filled), accuracy (no syntactic and spelling errors) and consistency (correct semantics and no logical errors). Completeness checks can be automated, while accuracy can only be checked for some fields (e.g., date format, well-formed URIs) and consistency can only be checked at a very limited extent (e.g., check if links can be resolved, check if MIME type of linked file is correct). We start from this work, and extend accuracy as follows. We consider the structuredness of value fields (e.g., dates, person names), as the lack of structure cannot or only with difficulty be reconstructed later if needed. Structuredness is defined as the degree to which each particle of a field represented by a simple data type can be directly accessed (this does not necessarily mean splitting the fields into subfields or attributes, this can also be achieved by providing well-defined parsing rules such as regular expressions). A further criterion is the use of controlled vocabularies, i.e., whether controlled vocabularies are used, whether they are publicly accessible and their level of quality (this is a property of a vocabulary that is assessed externally, and not for each record, e.g., using the eight classes of criteria proposed in [3]).

For assessing the mapping quality of metadata, the criteria completeness and consistency can be considered. The metadata formats used differ a lot among the various data providers, and require appropriate mappings. These mappings may not be lossless, but due to limitations of one of the formats, some loss of information or imprecision in mapping must be expected. The aim of mapping quality assessment is thus to quantify the loss of completeness and consistency of metadata documents resulting from mappings, in order to provide feedback to the experts defining the mapping, and to keep this loss as small as possible. Due to the scale of the problem expert assessment of mappings for a range of formats and a significant number of metadata documents is not feasible. Thus, automated methods to assess the quality of mappings are required, capable of measuring the loss in completeness and consistency w.r.t. expected imprecision or information loss due to the nature of the involved formats.

2 Assessment of Source Data Quality

In our approach we check the quality at three stages in the data flow through the EEXCESS framework. The first check is after calling the data provider API, the next after the transformation to the EEXCESS data model and the third after adding metadata with semantic enrichment. The proposed approach uses a set of checks, which are adapted to each data provider, using the knowledge about their respective data models. The basic set of checks can identify present metadata fields and check existence and validity of values in these metadata fields. The checks also handle multiple occurrences of the same metadata field in one record. Where applicable, validation against constraints of the data model is applied.

The EEXCESS project has defined a data model based on the Europeana Data Model (EDM)[3], which serves as the mapping target for aggregation. We have added to this data model the W3C PROV data model[4] to model the provenance of the metadata, especially in order to distinguish between the metadata which comes from the data provider and metadata added during semantic enrichment.

We have built a prototype[5] which uses data logged during calling the services from different data providers. Most services provide the data in XML, and if the service provides the data in JSON, the data are transformed to XML by an internal service of the prototype. In particular, we analyse the input data and records,

[3] http://pro.europeana.eu/page/edm-documentation

[4] http://www.w3.org/TR/prov-overview/

[5] Source code published at https://github.com/EEXCESS/data-quality.

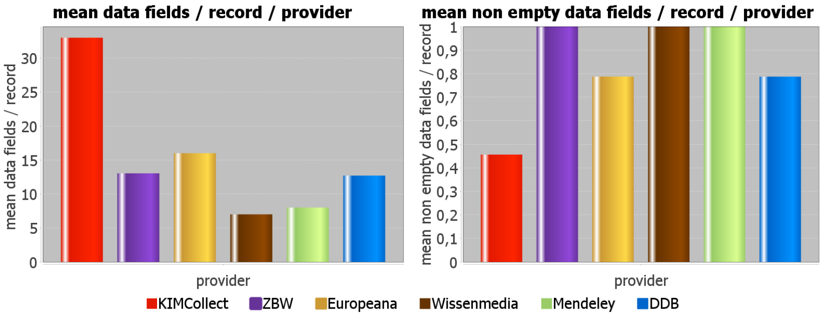

Fig. 1. Results of source metadata quality assessment: number of returned fields (left), fraction of non-empty fields (right).

the number of returned fields per record and empty and non-empty fields. As a result of the analysis the prototype generates statistics of the measured values and also generates charts.

For testing our prototype we use a randomly selected subset data containing over 6,000 records from six data providers. Some data providers include only metadata fields in their service response, if a value for the actual object is present. That is the reason why we calculate the mean value of submitted metadata fields per record for each data provider. The mean number of returned metadata fields varies between seven and 33 per record. In Figure 1 we show the graph of data fields per record per data provider and the relative number of non-empty fields.

We also analyse the use of controlled vocabularies of the data providers involved. Most of the used vocabularies are created by the data providers but are publicly accessible. Records from those data providers which use non-public vocabularies must be translated to other vocabularies during mapping.

3 Assessing Mapping Quality

The gold standard approach to assess the quality of a mapping is to compare a mapping result with a corresponding expert created reference. However, it is infeasible to implement this approach in an environment where a user configures and tests mappings while expecting immediate mapping quality feedback, and where the defined mappings are then applied on the fly without an option for intervention. Here, providing a ground truth mapping result by an expert without delay in the mapping creation process is impossible.

Our proposed approach for assessing the quality aims to meet the requirements of a tool-based mapping creation process without consulting an expert created ground truth. The mapping quality is assessed by performing a round trip mapping of a given metadata document and then detecting differences. The term round trip denotes that the documents to be compared for quality assessment are represented using the same metadata format. Thus as a precondition, bidirectional mappings for the underlying metadata format must be available. By

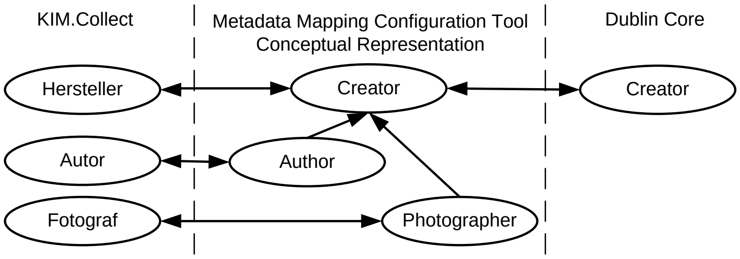

Fig. 2. Mapping paths between different metadata elements.

identifying the presence, absence, and representation of specific metadata properties in the mapping result with respect to the original document, statements about the mapping quality are made. For example, the absence of metadata properties in the mapping result indicates an impairment of the mapping quality and is a spot for possible improvements in the mapping specification. This approach is integrated in our metadata mapping configuration tool, which enables the creation of mappings between different XML-based metadata formats. The core part of the mapping configuration tool is an intermediate conceptual representation of metadata properties, which serves as a hub for mapping metadata between different formats. The metadata mapping configuration tool, which is available as a web application, and this conceptual representation are introduced in [4].

Two different variants of round trip mappings are supported by our metadata mapping configuration tool. The first variant considers only the internal intermediate conceptual representation of metadata properties, while the second variant also includes a specific target metadata format. The expected loss or imprecision between a pair of formats needs to be specified by an expert once per metadata format in the configuration tool. In Figure 2, the possible mapping paths between concepts of the two specific metadata formats and the conceptual metadata representation of the configuration tool are depicted. In case of assessing a round trip mapping via the conceptual representation there is no loss of information with respect to the involved metadata elements. Here, the KIM.Collect metadata elements `Autor` and `Fotograf` map to the conceptual metadata elements `Author` and `Photographer` and vice versa. In addition, there is a mapping path from these KIM.Collect metadata elements back to the element `Hersteller` (via `Creator`). This would lead to loss of information since `Hersteller` is not exactly the same as `Autor` or `Fotograf`. In case a lossless mapping option is available, a possible more general mapping will not be performed by the configuration tool. When considering the Dublin Core metadata elements for the round trip mapping, a loss of information has to be accepted. Here the round trip mapping from the elements `Autor` and `Fotograf` to the Dublin Core element `Creator` and back to the KIM.Collect element `Hersteller` is the only available option.

Finally, metadata elements in the original document and the mapping result are compared. For example, assume that a round trip mapping of a KIM.Collect metadata document including the metadata elements `Autor` and `Fotograf` based on the intended mapping paths, depicted in Figure 2, is performed. A round trip mapping using the intermediate representation leads to `Autor` and `Fotograf` elements in the resulting document. If the round trip mapping is performed via Dublin Core, only `Hersteller` elements will remain in the document. The presence of these elements indicates a correct mapping process, while their absence is a trigger to redefine the mapping instructions using the metadata mapping configuration tool.

4 Conclusion

In this paper, we have analysed the challenges of metadata quality assessment in a distributed system that aggregates metadata from a number of heterogeneous providers of cultural heritage objects on the fly. We use provenance metadata to track to origin of metadata fields and we have described the approaches we have implemented for assessing the source metadata quality and the quality of mappings between metadata formats, and we have presented our first results for those two problems.

Acknowledgements. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°600601, "EEXCESS".

References

1. Emanuele Bellini and Paolo Nesi. Metadata quality assessment tool for open access cultural heritage institutional repositories. In *Information Technologies for Performing Arts, Media Access, and Entertainment*, pages 90–103, 2013.
2. Thomas R. Bruce and Diane I. Hillmann. *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*, pages 238–256. ALA Editions, Chicago, 2004.
3. Evelyn Dröge. Criteria for vocabulary evaluation and comparison. Technical report, Humboldt-Universität zu Berlin, 2012.
4. Thomas Orgel, Martin Höffernig, Werner Bailer, and Silvia Russegger. A metadata model and mapping approach for facilitating access to heterogeneous cultural heritage assets. *International Journal on Digital Libraries*, 15(2-4):189–207, 2015.
5. B. Stvilia, L. Gasser, and M. Twidale. A framework for information quality assessment. *J. American Society for Information Science & Technology*, 58(12), 2007.